능동학습법을 이용한 한국어 대화체 문장의 효율적 의미 구조 분석

(Efficient Semantic Structure Analysis of Korean Dialogue Sentences using an Active Learning Method)

김 학 수 †

(Harksoo Kim)

요 약 목적 지향성 대화에서 화자의 의도는 화행과 개념열 쌍으로 구성되는 의미 구조로 근사화될 수 있다. 그러므로 지능형 대화 시스템을 구현하기 위해서는 의미 구조를 올바르게 파악하는 것이 매우 중요하다. 본 논문에서는 능동학습(active learning) 방법을 이용하여 효율적으로 의미 구조를 분석하는 모델을 제안한다. 제안 모델은 언어 분석에 따른 부담을 덜기위하여 형태소 자질들과 이전 의미 구조만을 입력 자질로 사용한다. 그리고 정확률 향상을 위하여 자연어 처리 분야에서 높은 성능을 보이고 있는 CRFs(Conditional Random Fields)를 기본 통계 모델로 사용한다. 일정 관리 영역에서 제안 모델을 실험한 결과는 기존 모델들과 비교하여 1/3 정도의 훈련데이타를 사용하고도 비슷한 정확률(화행 92.4%, 개념열 89.8%)을 나타내고 있음을 알 수 있었다.

키워드 : 능동학습, 의미 구조, 화행, 개념열

Abstract In a goal-oriented dialogue, speaker's intention can be approximated by a semantic structure that consists of a pair of a speech act and a concept sequence. Therefore, it is very important to correctly identify the semantic structure of an utterance for implementing an intelligent dialogue system. In this paper, we propose a model to efficiently analyze the semantic structures based on an active learning method. To reduce the burdens of high-level linguistic analysis, the proposed model only uses morphological features and previous semantic structures as input features. To improve the precisions of semantic structure analysis, the proposed model adopts CRFs(Conditional Random Fields), which show high performances in natural language processing, as an underlying statistical model. In the experiments in a schedule arrangement domain, we found that the proposed model shows similar performances (92.4% in speech act analysis and 89.8% in concept sequence analysis) to the previous models although it uses about a third of training data.

Key words : active learning, semantic structure, speech act, concept sequence

1. 서 론

대화시스템은 인간과 컴퓨터가 자연스럽게 의사소통할 수 있도록 인간의 언어를 이해하고 그에 적합한 응답을 찾아 제시해 주는 지능형 컴퓨터 프로그램을 말한다. 이러한 대화시스템을 구현하기 위해서는 대화를 구성하는 각 발화에 숨어있는 화자의 의도를 정확히 분석해야 한다. 본 논문에서는 화자의 의도를 화행(speech act)과 개념열(concept sequence)의 쌍으로 이루어진 의미 구조로 일반화한다. 표 1은 본 논문에서 정의한 의미구조의 예를 보여준다. 표 1에서 보는 것과 같이 화행은 발화에 표현된 영역 독립적인 일반 의도를 나타내며, 개념열은 영역 종속적인 세부 의미를 나타낸다.

· 이 연구(논문)는 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다. 또한 부분적으로 강원대학교 정보통신연구소의 지원을 받았습니다.

† 정 회 원 : 강원대학교 컴퓨터정보통신공학 교수
nlpdrkim@kangwon.ac.kr

논문접수 : 2007년 3월 8일
심사완료 : 2008년 3월 31일

표 1 의미 구조의 예

발화	의미 구조(화행, 개념열)
(1) 내일 일정 좀 알려줘	request, timetable-select
(2) 내일 오후 1시에 홍길동님과 공대6호관 501호에서 미팅이 있습니다.	response, timetable-select
(3) 시간이 바뀌었어	inform, timetable-update
(4) 언제로 바뀌었습니까?	ask-ref, timetable-update-time
(5) 오후 3시	response, timetable-update-time
(6) 약속시간을 오후 1시에서 3시로 변경하였습니다.	inform, timetable-update-time

화행과 개념열의 특징은 두 가지 모두 문맥에 의존하기 때문에 표층적인 발화 형태만으로 추론하는 것이 매우 어렵다는 것이다[1]. 예를 들어, 표 1의 발화 (5)를 표층적으로만 분석하면 다음과 같은 두 가지의 의미 구조가 가능하다.

- 'inform & timetable-select-time': 현재 설정되어 있는 약속 시간을 알려주는 행위
- 'response & timetable-update-time': 약속 시간 변경을 요청하는 응답 행위

이러한 모호성을 해결하기 위해서는 발화 (5)의 문맥을 고려해야 한다. 이 예제의 경우에 바로 이전의 발화 (4)를 고려하면 올바른 의미 구조인 'response & timetable-update-time'을 선택할 수 있다.

기존의 연구들은 계획 추론 모델(plan inference model)을 위한 레서피(recipe)나 영역 의존적인 지식에 기초해 왔다[2-4]. 그러나 영역 의존적인 지식을 이용한 연구들은 응용 영역을 확장하거나 변경할 때마다 대용량의 지식들을 수동으로 다시 구축해야 한다는 단점이 있다. 최근에는 이러한 단점을 극복하기 위하여 통계 기반의 다양한 기계 학습 모델들이 이용되고 있다[5-7]. 통계 기반의 기계 학습 모델들은 대용량의 학습 데이타로부터 자동으로 추출된 정보를 이용하여 의도 범주에 사용자의 발화들을 효과적으로 할당한다. 그러나 기존의 통계 기반의 모델을 이용하여 높은 정확률을 얻기 위해서는 많은 시간과 노력이 드는 대용량의 학습 데이타를 구축해야 한다. 이러한 문제점을 해결하기 위해서 본 논문에서는 능동학습(active learning) 방법을 이용하여 적은 양의 학습 데이타만으로도 높은 정확률을 보이는 효율적인 의미 구조 분석 모델을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 능동학습 기반의 효율적인 의미 구조 분석 모델을 제안한다. 3장에서는 실험 데이타를 설명하고 실험 결과를 분석한다. 마지막으로 4장에서 결론을 내리고 향후과제를 제시한다.

2. 능동학습을 이용한 의미 구조 분석 시스템

2.1 의미 구조의 정의

의미 구조 분석에 관한 기존의 연구들은 합의된 표준이 없이 영역에 따라 매우 주관적으로 화행을 정의하고 사용해 왔다. 화행을 영역 의존적으로 정의하여 사용하면 해당 영역에서 매우 정밀하게 사용자의 의도를 분석할 수 있다는 장점이 있지만, 영역을 확장하거나 변경할 경우에 화행 수를 늘리거나 재정의해야 한다는 단점이 있다. 이러한 문제를 줄이기 위해서 본 논문에서는 이현정(1997)[8]을 기초로 영역 독립적이며 포괄적인 화행을 정의하고, 영역 의존적인 의미들은 개념열을 통하여 보충하는 방법을 사용한다. 표 2는 본 논문에서 사용하는 화행을 보여준다.

목적 지향 대화는 상대방으로부터 특정 정보를 얻거나 전달하기 위해서 진행되며, 기존의 많은 대화 시스템들은 데이타베이스 연산을 이용하여 이러한 현상들을 모델링하였다[9,10]. 이러한 기존 연구들에 기초하여 본 논문에서는 표 3과 같이 일정관리 영역에 필요한 2가지 테이블, 4가지 연산자, 8가지 필드를 3층 구조 개념열 부착 방법[11,12]에 따라 64가지의 개념열을 정의하여 사용한다.

2.2 의미 구조 분석 모델

하나의 대화를 구성하는 첫 번째 발화 U_1부터 i번째 발화 U_i까지의 의미 구조 $SS_{1,i}$를 계산하는 확률 모델에서 화행과 개념열이 서로 독립이라고 가정하면 $P(SS_{1,i}|U_{1,i})$

표 2 영역 독립 화행

화행	설명	화행	설명
Greeting	대화 서두의 인사말	Request	행위를 요청
Expressive	대화 후미의 인사말	Ask-confirm	이전 발화의 확인
Opening	실제 대화의 시작	Confirm	확인 발화의 응답
Ask-ref	WH-질문	Inform	정보 제공
Ask-if	YN-질문	Accept	호응
Response	응답		

표 3 일정관리 영역의 개념열

테이블명	연산자	필드명
timetable alarmtable	insert delete select update	agent date day-of-week time person place content field

는 식 (1)과 같다.

$$P(SS_{1,i}|U_{1,i}) \approx P(SA_{1,i}|U_{1,i})P(CS_{1,i}|U_{1,i}) \quad (1)$$

화자는 개인의 언어적 습관에 따라 동일한 의미의 문장을 단어의 순서를 바꾸거나 불필요한 단어를 생략하는 방법 등을 통하여 다양한 형태로 표현하므로 $P(SA_{1,i}|U_{1,i})$나 $P(CS_{1,i}|U_{1,i})$를 직접 계산하는 것은 매우 힘들다. 이러한 문제를 해결하기 위해서 본 논문에서는 발화가 문장 자질들의 집합에 의해서 일반화될 수 있으며, 현재 발화의 문장 자질들이 이전 발화나 이후 발화의 의미 구조에 영향을 미치지 않는다고 가정하고, 식 (1)을 식 (2)와 같이 고쳐 쓴다.

$$\begin{aligned} P(SS_{1,i}|U_{1,i}) &\approx P(SA_{1,i}|F_{1,i})P(SA_{1,i})P(CS_{1,i}|F_{1,i})P(CS_{1,i}) \\ &\approx P(SA_i|F_i)P(SA_{1,i})P(CS_i|F_i)P(CS_{1,i}) \end{aligned} \quad (2)$$

식 (2)에서 F_i는 i번째 발화의 문장 자질 집합으로써 어휘 자질과 품사 자질로 구성된다[13]. 어휘 자질은 품사가 부착된 어휘를 말하며, 품사 자질은 품사 바이그램(bigram)을 말한다. 예를 들어, '안녕하세요?'라는 발화의 형태소 분석 결과가 '안녕/ncp+하/xsp+세요/ef+ ?/sf'라고 했을 때, 어휘 자질은 '안녕/ncp, 하/xsp, 세요/ef, ?/sf'이고, 품사 자질은 'ncp-xsp, xsp-ef, ef-sf'이다.

어휘 자질과 품사 자질의 추출이 끝나면 식 (3)과 같이 각 자질과 범주(화행 또는 개념열) 사이의 χ^2 통계량을 계산하고, 값이 큰 상위 n개의 자질을 선택하여 정보량이 적은 자질들을 제거한다[12,14].

$$\chi^2(f,c) = \frac{(A+B+C+D)\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (3)$$

식 (3)에서 f는 자질(어휘 자질 또는 품사 자질)을 의미하며, c는 화행 또는 개념열 범주를 의미한다. A는 c에 속해 있는 발화 중에서 f를 포함하는 발화 수, B는 c 이외의 범주에 속해 있는 발화 중에서 f를 포함하는 발화 수, C는 c에 속해 있는 발화 중에서 f를 포함하지 않는 발화 수, D는 c 이외의 범주에 속해 있는 발화 중에서 f를 포함하지 않는 발화 수를 의미한다.

화행과 개념열은 문맥에 의존하지만 이전의 모든 대화 기록을 고려하여 현재의 화행과 개념열을 확률적으로 결정하는 것은 현실적으로 불가능하다. 본 논문에서는 1차 마코프(Markov) 가정을 통하여 식 (2)를 식 (4)와 같이 고쳐 쓴다.

$$\begin{aligned} P(SS_i|U_i) \approx & P(SA_i|F_i)P(SA_i|SA_{i-1}) \\ & P(CS_i|F_i)P(CS_i|CS_{i-1}) \end{aligned} \quad (4)$$

그리고 화행 분석 확률 $P(SA_i|F_i)P(SA_i|SA_{i-1})$과 개념열 분석 확률 $P(CS_i|F_i)P(CS_i|CS_{i-1})$을 식 (5)와 같은 CRFs(Conditional Random Fields)를 이용하여 계산한다.

$$\begin{aligned} P_{CRF}(SA_{1,n}|F_{1,n}) &= \frac{1}{Z(F_{1,n})} exp\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(SA_{i-1}, SA_i, F_{1,n}, i)\right) \\ P_{CRF}(CS_{1,n}|F_{1,n}) &= \frac{1}{Z(F_{1,n})} exp\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(CS_{i-1}, CS_i, F_{1,n}, i)\right) \end{aligned} \quad (5)$$

그림 1은 본 논문에서 제안한 CRFs 기반의 의미 구조 분석 모델을 확률 그래프 형태로 표현한 것이다.

그림 1에서 보는 것과 같이 CRFs는 다양한 입력 노드의 값이 주어졌을 때 지정된 출력 노드의 조건부 확률값을 계산하기 위한 무방향성 그래프 모델이다. CRFs는 HMM(Hidden Markov Model)의 단점인 독립 가정을 완화시키는 효과가 있으며, MEMM(Maximum Entropy Markov Model)의 단점인 레이블 편향 문제(label bias problem)를 극복할 수 있다는 장점을 가지고 있어서 최근 자연어처리 분야에서 가장 많이 사용되는 통계기반의 기계 학습 모델이다. 식 (5)에서 $SA_{1,n}$과 $CS_{1,n}$은 n개의 발화로 구성된 대화에서 입력된 문장 자질 $F_{1,n}$에 대한 출력 화행과 개념열을 의미한다. $Z(F)$는 정규화 요소이며, $f_k(SA_{i-1}, SA_i, F_{1,n}, i)$와 $f_k(CS_{i-1}, CS_i, F_{1,n}, i)$는 화행 분석과 개념열 분석을 위한 자질 함수로서 출력열 ($<SA_{i-1}, SA_i>$와 $<CS_{i-1}, CS_i>$) 에 대해서 i번째 발

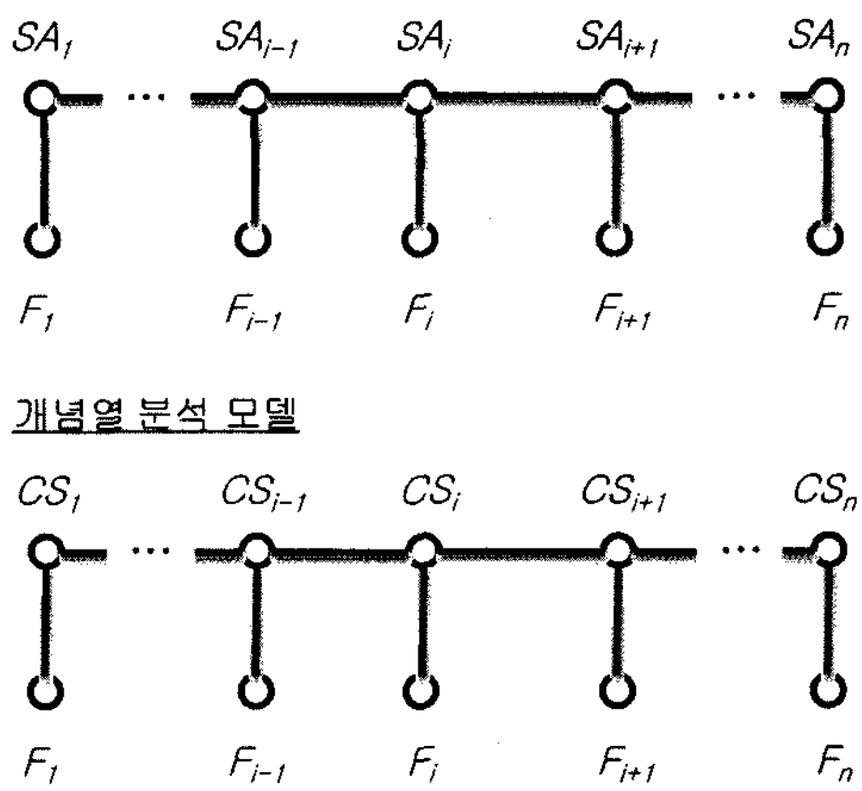


그림 1 CRFs 기반의 의미 구조 분석 모델의 그래프 구조

화의 문장 자질 집합 F_i에서 자질 f_k가 나타나면 1을, 그렇지 않으면 0의 값을 가진다. λ_k는 자질 f_k의 가중치이다.

2.3 능동학습 방법

그림 2는 능동학습을 이용한 의미 구조 분석 시스템의 구조도이다. 그림 2에서 보듯이 제안 시스템은 학습 시스템과 적용 시스템으로 구성된다. 두 개의 서브시스템(sub-system)은 동일한 방법을 이용하여 입력 문장으로부터 자질을 추출하고 정보량이 큰 것들을 선택한다. 학습 시스템은 CRFs(Conditional Random Fields)[15,16]를 이용하여 확률 모델을 학습한다. 먼저, 현재 보유한 의미 구조 부착 말뭉치를 이용하여 확률 모델을 완성한다. 그리고 학습에 참여하지 않은 데이타를 이용하여 의미 구조 분석을 수행한 후, 신뢰도가 낮은 하위 k개의 문장들을 선별한다. 마지막으로 선별된 k개의 문장에 수동으로 의미 구조를 부착하고, 학습 과정을 반복하면서 확률 모델의 성능을 향상시킨다. 적용 시스템은 최종 완성된 CRFs 확률 모델을 기반으로 하여 사용자가 입력한 대화 문장의 의미 구조를 분석한다.

통계 기반의 기계 학습 모델을 이용하여 높은 정확률을 얻기 위해서는 대용량의 의미 구조 부착 데이타를 구축해야 한다. 능동학습은 최소한의 학습 데이타를 이용하여 학습 효과가 높을 것으로 추정되는 데이타를 선정하고 추가하는 작업을 반복적으로 수행하여 학습효과를 높이는 방법이다[17,18]. 능동학습은 학습단계와 신뢰도 측정단계로 나뉘어 진다. 학습단계에서는 현재 보유한 학습 데이타 집합에 기본 학습 모델(본 논문에서는 CRFs 모델)을 적용한다. 신뢰도 측정단계에서는 학습에 참여하지 않은 데이타에 대한 평가를 수행하고, 신뢰도가 낮은 데이타에 대하여 전문가가 수동으로 정답을 부착한다. 전문가에 의해 정답이 부착된 신규 학습 데이타들은 기존의 학습 데이타 집합에 추가된다. 그림 3은 의미 구조 분석에 능동학습 방법을 적용하는 방법을 보여준다.

3. 실험 및 평가

3.1 실험 데이타

실험을 위하여 일정관리 영역에서 Wizard-Of-Oz 방식으로 시뮬레이션(simulation)한 대화 말뭉치를 수집한 후, 수동으로 화행과 개념열을 부착하였다. 수집된 말뭉치는 일정 추가, 삭제, 변경과 관련된 내용을 포함하고 있다. 말뭉치의 구성은 956개의 대화(21,336개의 발화)로 구성되며, 대화당 평균 발화의 수는 22.32개이다. 능동학습 데이타와 성능평가 데이타의 비율은 4:1로 나누

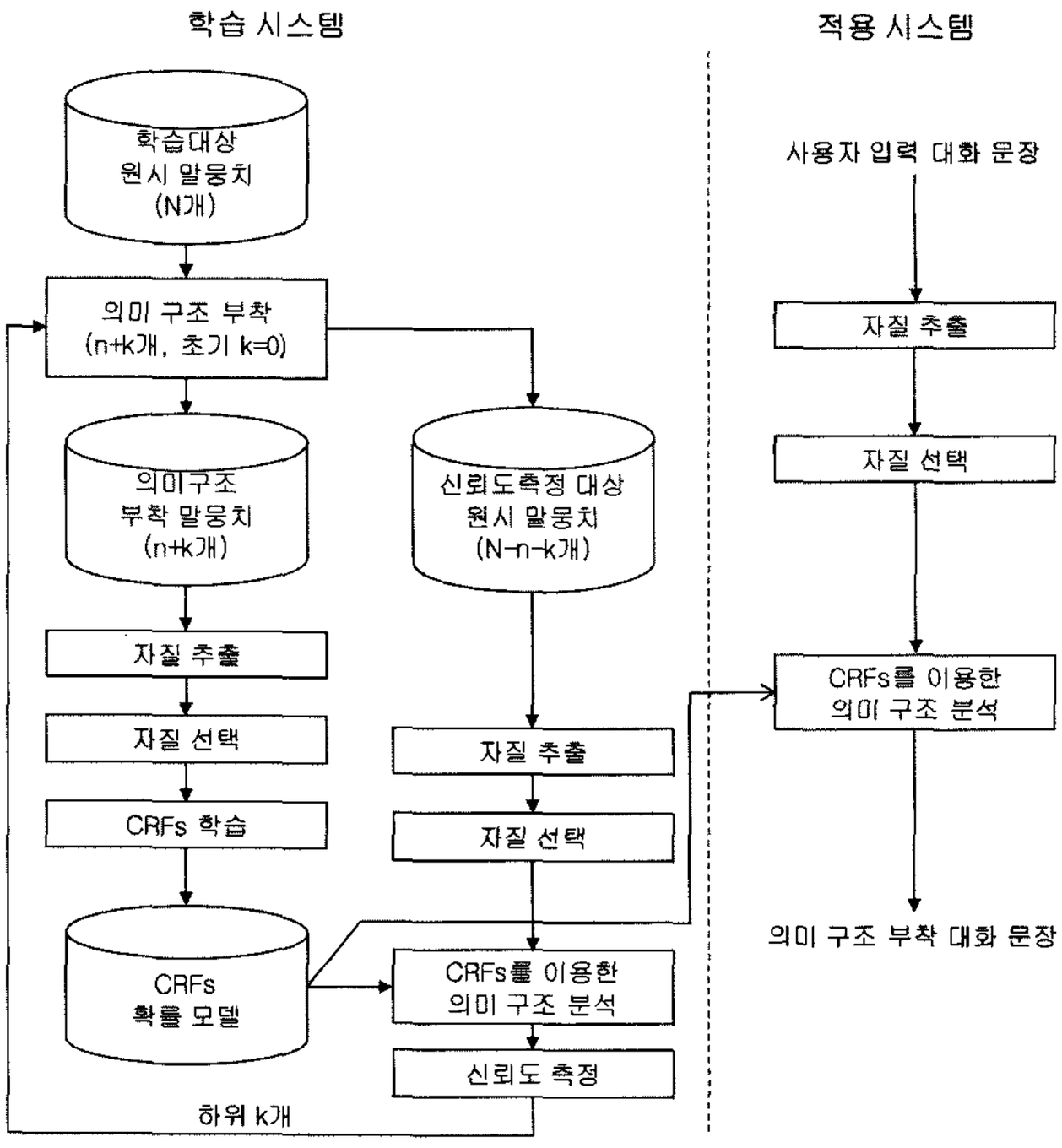


그림 2 의미 구조 분석 시스템의 구조도

1. 전체 데이터를 능동학습 데이터(80%)와 성능평가 데이터(20%)로 구분
2. 능동학습 데이터의 0.6%에 대하여 수동으로 의미 구조 부착 수행하고 훈련데이터로 사용
3. CRFs를 이용한 통계 모델 학습
4. 의미 구조가 부착되지 않은 데이터를 이용하여 통계 모델 평가
5. 신뢰도(확률값)가 낮게 나오는 하위 n개의 문장을 선택 (화행, 개념열 각각에 대해서 각각 n개 선택)
6. 화행과 개념열 모두에 해당하는 k개의 문장에 대하여 수동으로 의미 구조 부착 수행
7. 0.6%+k개의 훈련데이터로 사용하여 하위 신뢰도의 변화가 없거나 일정 수준 이상 나올 때까지 (2)에서 (7)을 반복
8. 성능평가 데이터를 이용하여 최종 정확률 평가

그림 3 의미 구조 분석을 위한 능동학습 방법

었으며, 문장 자질의 수는 800개로 제한하였다. CRFs 모델의 내부 설정 인자로 추정알고리즘은 L-BFGS를 이용하였으며[19], 희소 데이타 문제를 위한 평탄화 요소는 Gaussian Prior를 이용하였다[15]. Gaussian Prior의 값은 10으로 설정하였으며, 훈련 반복 회수는 30으로 설정하였다.

3.2 실험 결과

표 4는 0.6%의 초기 훈련데이타를 시작으로 능동학습을 진행하면서 매회 마다 성능평가 데이타를 이용하여 정확률을 측정한 결과를 보여준다.

표 4를 분석한 결과, 신뢰도가 낮은 하위 문장들에서 화행과 개념열 분석 모두에 포함되는 것이 초기에는 적었지만 능동학습을 진행하면서 점점 많아짐을 관찰할 수 있었다. 이것은 화행 분석과 개념열 분석 문제가 서로 독립적인 성격이 크지만 완전 독립은 아니라는 사실을 말해주는 것으로 보인다. 의미 구조 분석에 실패한 경우들을 살펴본 결과도 비슷한 결론에 도달할 수 있었다. 화행 분석에 실패한 문장에서 개념열을 문맥 정보로 사용해야만 하는 경우가 있었으며, 개념열 분석에 실패한 문장에서 화행을 문맥 정보로 사용해야 하는 경우를 발견할 수 있었다. 이것 또한 화행과 개념열이 서로 완전 독립이 아니라는 것을 뒷받침해준다고 할 수 있다. 표 5는 화행 분석 시에 개념열을 고려해야하는 예를 보여준다.

표 4 능동학습에 따른 정확률 변화

훈련데이타(%)	화행 정확률(%)	개념열 정확률(%)
0.6	76.2	56.9
4.7	79.5	69.3
9.2	83.8	78.2
15.7	87.1	83.3
22.9	90.7	86.2
27.7	92.4	89.2
30.6	92.4	89.8

표 5에서 발화 (3)의 화행을 결정하기 위하여 발화 (2)의 화행만을 고려한다면 발화 (3)의 화행은 inform이 될 가능성이 매우 크다. 그러나 만약 발화 (3)의 개념열을 고려한다면 발화 (3)이 발화 (2)와 매우 밀접하게 연관되어 있고 그것의 의미는 변경된 날짜라는 것을 추측할 수가 있다. 이러한 사실에 기초하면 발화 (3)의 화행은 변경된 시간을 시스템에게 전달하기 위한 response라는 것을 알 수 있다. 이러한 문제를 근본적으로 해결하기 위해서는 식 (1)을 수정한 새로운 통계 모델의 정립이 필요할 것으로 생각된다.

표 6은 능동학습을 사용하지 않고 훈련데이타로부터 0.6%, 15.7%, 30.6%의 말뭉치를 무작위로 추출하여 학습한 후, 성능평가 데이타를 이용하여 정확률을 측정한 결과를 보여준다. 표 6에서 보는 것과 같이 능동학습을 이용한 경우가 일반 학습을 이용한 경우보다 화행 분석과 개념열 분석 모두에서 높은 정확률을 보였다. 특히 훈련데이타의 15.7%를 사용한 것보다 30.6%를 사용한

표 6 일반적인 학습에 따른 정확률 변화

훈련데이타(%)	화행 정확률(%)	개념열 정확률(%)
0.6	76.2	56.9
15.7	86.5	78.0
30.6	87.5	82.1

표 5 화행 분석 시에 개념열을 고려해야 하는 경우의 예

발화	화행	개념열
(1) 시스템: 무엇이 바뀌었습니까?	Ask-ref	Timetable-update
(2) 사용자: 약속 날짜가 바뀌었어.	Response	Timetable-update-date
(3) 사용자: 3월 20일이야.	Response	Timetable-update-date

표 7 기존 모델과의 정확률 비교

모델(훈련데이타 %)	화행 정확률(%)	개념열 정확률(%)
En-2005(100)	90.4	-
Lee-2006(100)	91.9	89.9
제안 모델(30.6)	92.4	89.8

것에서 더 큰 정확률 차이를 보였다. 이것은 능동학습을 이용한 제안 모델이 매우 효율적임을 보여주는 것이라고 할 수 있다.

표 7은 훈련데이타 모두를 사용하는 기존의 두 모델(En-2005[20], Lee-2006[12])과 24.5%만을 사용하는 제안 모델 사이의 정확률을 비교한 것이다.

표 7에서 En-2005는 제안 모델과 동일한 입력 자질(문장 자질과 문맥 자질)을 SVM(Support Vector Machine)의 입력으로 사용하여 화행을 분류하는 모델이다. (은종민, 2005)[20]에 따르면 문장 자질과 문맥 자질에 구문 자질(구문분석기를 이용하여 반자동으로 구축되는 문법적 역할과 관련된 자질)을 추가하였을 경우에 화행 분류 성능이 향상되는 것으로 알려져 있다. 그러나 본 논문에서는 (은종민, 2005)에서 사용한 것과 동일한 구문분석기를 구할 수 없어서 구문자질이 추가된 실험은 수행할 수 없었다. Lee-2006은 제안 모델과 유사한 입력 자질을 기반으로 하면서 현재 발화의 개념열 분류 결과를 화행 분류에 이용하여 정확률을 향상시킨 신경망 모델이다. 표 7에서 보듯이 능동학습을 사용하는 제안 모델은 기존 모델의 1/3 수준인 30.6%의 학습 데이타를 사용하고도 비슷한 수준의 정확률을 얻을 수 있었다. 이러한 사실에 기초하여 제안 모델이 기존의 모델들보다 훨씬 효율적임을 알 수 있었다.

4. 결 론

본 논문에서는 능동학습을 이용하여 화자의 의도를 효율적으로 파악하는 새로운 의미 구조 분석 모델을 제안하였다. 의미 구조 분석에 능동학습을 적용한 모델과 그렇지 않은 경우를 비교하여 볼 때 1/3 수준의 훈련데이타만을 이용하고도 비슷한 수준의 정확률을 얻을 수 있었다. 또한 다른 자연어처리 분야에서 높은 성능을 보이고 있는 CRFs를 능동학습과 접목시켜 의미 구조 분석 분야에서의 적용 가능성을 알아보았다.

향후 과제로는 첫째, 제안된 방법이 일반적으로 적용될 수 있는지를 보기 위하여 일정관리 영역이 아닌 다른 영역들에 대한 다양한 실험이 필요할 것으로 보인다. 둘째, 능동학습시에 신뢰도가 낮은 문장을 자동으로 선택할 수 있는 객관적이고 과학적인 방법에 대한 연구가 필요할 것으로 보인다.

참 고 문 헌

[1] Levin, L., Langley, C., Lavie, A., Gates, D. Wallace, D., and Peterson, K., "Domain specific speech acts for spoken language translation," in *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue*, 2003.

[2] Caberry, S., *A pragmatics-based approach to ellipsis resolution*. Computational Linguistics, Vol.15, No.2, pp. 75-96, 1989.

[3] Lambert, L. and Caberry, S., "A tripartite plan-based model of dialogue," in *Proceedings of ACL 1991*, pp. 47-54, 1991.

[4] Litman, D. J. and Allen, J. F., *A plan recognition model for subdialogues in conversations*, Cognitive Science, Vol.11, pp. 163-200, 1987.

[5] Langley, C., "Analysis for speech translation using grammar-based parsing and automatic classification," in *Proceedings of the ACL Student Research Workshop*, 2002.

[6] Lee, S. and Seo, J., "An analysis of Korean speech act using hidden Markov model with decision trees," in *Proceedings of ICCPOL 2001*, pp. 397-400, 2001.

[7] Samuel, K., Caberry, S., and Vijay-Shanker, K., "Computing dialogue acts from features with transform-based learning," in *Proceedings of the AAAI Spring Symposium*, pp. 90-97, 1998.

[8] 이현정, *한국어 대화체 문장의 화행 분석*. 석사학위논문, 서강대학교, 1997.

[9] Goddeau, D., Meng, H., Polifroni, J., Seneff, S., and Busayapongchai, S., "A form-based dialogue manager for spoken language applications," in *Proceedings of International Conference on Spoken Language Processing*, pp. 701-704, 1996.

[10] Kim, H., Seon, C., and Seo, J., *A dialogue-based information retrieval assistant using shallow NLP techniques in online sales domains*, IEICE Information and Systems, Vol.E88D, No.5. pp. 801-808, 2005.

[11] Kim, H., *A dialogue-based NLIDB system in a schedule management domain*. Lecture Notes in Computer Science, Vol.4362, pp. 869-877, 2007.

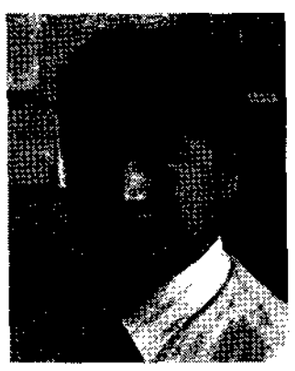
[12] Lee, H., Kim, H., and Seo, J., *Efficient domain action classification using neural networks*. Lecture Notes in Computer Science, Vol.4233, pp. 150-158, 2006.

[13] 김경선, 서정연, *자질 선택 기법을 이용한 한국어 화행 결정*, 한국정보과학회 논문지, 제30권 3호, pp. 278-284, 2003.

[14] Yang, Y. and Pedersen, J. O., "A comparative study on feature selection in text categorization," in *Proceedings of ICML 1997*, 1997.

[15] Lafferty, J., McCallum, A., and Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Pro-*

ceedings of ICML 2001, 2001.
[16] Pinto, D., McCallum, A., Wei, X., and Croft, W. B., "Table extraction using conditional random fields," in *Proceedings of SIGIR 2003*, 2003.
[17] Cohn, D., Ghahramani, Z., Jordan, M., *Active learning with statistical models*. Journal of Artificial Intelligence Research, Vol.4, pp. 129-145, 1999.
[18] Riccardi, G. and Hankkani-Tur, D., *Active learning: theory and applications to automatic speech recognition*, IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp. 504-511, 2005.
[19] Fei, S. and Pereira, F., "Shallow parsing with conditional random fields," in *Proceedings of HLT & NAACL 2003*, 2003.
[20] 은종민, 이성욱, 서정연, *지지벡터기계(support vector machines)를 이용한 한국어 화행분석*, 한국정보처리학회 논문지, 제12B권 3호, pp. 365-368, 2005.



김 학 수
1996년 건국대학교 전자계산학과 학사
1998년 서강대학교 컴퓨터학과 석사
2003년 서강대학교 컴퓨터학과 박사
2004년 University of Massachusets, Amherst 박사후연구원. 2005년 한국전자통신연구원 선임연구원. 2006년~현재 강원대학교 컴퓨터정보통신공학전공 조교수. 관심분야는 한국어정보처리, 생략 및 대용어 처리, 대화 인터페이스 시스템, 정보검색 시스템, 질의응답 시스템